Analyzing Census Bureau Data In SAS® Studio

Transcripts

Table of Contents

Introduction to Analyzing Census Bureau Data in SAS Studio | Video 1 of 6 | 9:30 minutes 3
Access Census Bureau Data | Video 2 of 6 |12:25 minutes 6
Import Census Bureau Data into SAS Studio | Video 3 of 6 | 13:41 minutes..... 10
Visualize Census Bureau Data in SAS Studio | Video 4 of 6 | 17:19 minutes 14
Prepare Census Bureau Data in SAS Studio | Video 5 of 6 | 16:45 minutes 19
Analyze Census Bureau Data in SAS Studio | Video 6 of 6 | 20:40 minutes 24

Introduction to Analyzing Census Bureau Data in SAS Studio | Video 1 of 6 | 9:30 minutes

Hi, everyone. My name is Luna Bozeman, and I'm a technical trainer at SAS. Welcome to your tutorial on Analyzing Census Data in SAS Studio. This tutorial series contains six videos, the first being this overview video. In subsequent tutorials, I'll walk you through the process of accessing US census data and exploring it using SAS Studio. You'll learn to import, visualize, prepare, and analyze data using SAS.

Let's start with the data. Every year, the US Census Bureau publishes demographic, socioeconomic, housing, and business statistics for all communities in the country. Whether you're a market analyst wanting to learn more about your customers or a student working on a case study to learn more about your community, the Census Bureau makes it easy for you to access the data you need to gain those insights.

As the industry leader in analytics, SAS has many tools that you can use to analyze data. The tool that we'll use throughout this series is SAS Studio, a browser-based SAS programming interface. But you don't need to know how to write SAS code to use SAS Studio because it also contains visual point-and-click tasks that enable you to take advantage of the power of SAS.

In this tutorial series, you'll learn how to use these tasks to analyze census data. If you already have access to SAS Studio, that's great. If you'd like to use your own installation to follow along with the tutorials and complete the practices, you need to make sure you have three things-- SAS 9.4 maintenance release 3 or later, a license for SAS/ACCESS Interface to PC Files, and the SAS/STAT product.

If you're not sure what you have, check the notes for this tutorial below or on the main tutorial page. If you have those requirements, you can skip ahead to the Setting Up the Tutorial Data chapter in this video. If you don't have those or you don't have access to SAS Studio, don't worry. Follow the instructions in the Accessing SAS Studio Using SAS OnDemand for Academics chapter in this video to access SAS Studio. It's free for learners.

Let me show you how easy it is to create a SAS OnDemand for Academics account. If you would like a written version of the instructions that contains the links that I mentioned, look in the description below or on the main tutorial page. In your browser, go to welcome.oda.sas.com If you already have a SAS profile, then great. You can enter in your username and password to begin the registration process. But if you don't have a SAS profile, you can click on where it says "Don't have a SAS profile?" and follow through the instructions to create your SAS profile.

Once you have a SAS profile, again, come back to welcome.oda.sas.com and then sign in using your SAS profile. The first time you sign in, you will be prompted to select a home region. Once your sign up request process has been completed, you'll receive an email letting you know that you're ready to start using SAS OnDemand for Academics. Simply return to welcome.oda.sas.com, sign in, and then under Applications select SAS Studio. Now you're ready to use SAS Studio.

Whether you're using your own installation of SAS Studio or accessing it through SAS OnDemand for Academics, you need to set up the tutorial data if you want to follow along with the videos and complete the practices. Let me walk you through the steps. You can always find these instructions in the description below or on the main tutorial page.

Before you begin to work in SAS Studio, you need to download the tutorial materials. You can find a link to the tutorial materials in the description below or on the main tutorial page. Once you have the materials downloaded, come back in to SAS Studio. Taking a look at SAS Studio, you will see by default the Navigation pane on the left-hand side and then the work area on the right-hand side.

In the Navigation pane, you'll see that the Server Files and Folder section is expanded by default. In this section is where we're going to have to create a folder to store all of our work throughout this tutorial series. So what I'll do is I'll expand Files and then navigate to a location of your choice. I want to create my folder in the Home directory, so I'll right click where it says Files Home, then select New, then Folder. I'll name this folder Census Data Analysis and click Save.

Once you have that folder ready to go, we next need to upload the createdata.sas program to this tutorial folder, and then run the program to create the data that we need for this tutorial series. This program is included in the tutorial materials you downloaded. To upload the SAS program into my SAS environment, I'll select the Census Data Analysis folder and then click on the Upload icon.

Next up, click Choose Files and then I'll navigate to my createdata.sas program. Select it, click Open, and then select Upload. So now whenever I expand my Census Data Analysis folder, I'll see that createdata.sas program available. I'm going to go ahead and double click on that program to open it up in a new tab in the work area.

Now, before running this program, we need to specify the path to that Census Data Analysis folder that we just created. The easiest way to retrieve that path is to right click on the Census Data Analysis folder and then select Properties. Go ahead and highlight the location box, select Control C on your keyboard, and then select Close. Now with that path copied, in the program, I'm going to come to the line that says %LET path=insertpath; all you need to do is highlight where it says insertpath, and then go ahead and select Control V on your keyboard to paste in that path that we just copied. That is the only change that you need to make to this program, so go ahead and click Run. There's a little running man icon.

So on the generated report on the Results tab, you should see that nine tables are listed in the report. If you go and actually take a look at your Census Data Analysis folder, you'll see those nine tables listed in a folder along with three Excel files. You'll see educationattainment_s, medianhomevalue_s, and then totalpopulation_s there.

When you use point-and-click tasks in SAS Studio, you often need to specify an input SAS table to use, or an output SAS table to save the output information. The location of these tables often needs to be specified in the form of a library. Now, a library is simply just a shortcut or a pointer to a collection of SAS tables that are in a particular location.

So what I'm going to do is in order to create an access SAS tables in the Census Data Analysis folder, I'm going to create a library pointing to that folder. The easiest way to do this is to right click the Census Data Analysis folder, select Create, and select Library. I'll name the library Census and I'm also going to check this checkbox that says, "Re-create this library at start-up." Now, what usually happens is with user-defined libraries, whenever you close out of SAS Studio, those user-defined libraries are deleted. So to make sure that I always have the Census library available to me, I just need to select this checkbox so that it's always getting re-created. All right. I'll click OK.

You now have access to SAS Studio and the tutorial data, so you're almost ready to start the video series. One last thing I want to mention is the tutorial notes. The tutorial notes contain not only the steps you'll see in the videos, but also practices for you to try out what you've learned. There are even challenge practices that will go beyond what you'll see in the videos so that you can learn about other techniques you can use to analyze census data. You can find the PDF tutorial notes in the tutorial materials that you downloaded or on the main tutorial page.

Now you're ready to follow along with the Analyzing Census Data in SAS Studio tutorial series. You have all the data you need to jump into any video you're interested in. But for the best experience, I recommend watching the videos in order. Let's get started.

Access Census Bureau Data | Video 2 of 6 |12:25 minutes

Hi, everyone. Welcome to the Analyzing Census Data in SAS Studio tutorial series. In this video, I'll show you how you can access and download data that the US Census Bureau makes available. Then you'll learn to prepare this data for import into SAS Studio. You'll see how to import data in the next video. If you want to follow along with me, make sure you watch the introduction video to set up your environment and tutorial data. You can find the link to it below or on the main editorial page. Let's jump right in.

So the first thing I'm going to do is, in a browser, I'm going to go to data.census.gov. So data.census.gov is going to provide a centralized platform to access demographic and economic data from the United States Census Bureau. Now, this is actually not the only interactive tool that the Census Bureau makes available to access data, but it's the one that we'll be using throughout this tutorial series.

Once you get to data.census.gov on this main landing page, there's two main ways you can go about looking for the data that you're interested in. You can use the free form single search bar. And you can do this if you are just wanting to look for a quick statistic or you're looking for a profile for a single geography. Or you can alternatively take advantage of advanced search. And this is more for your complex searches, like a particular survey, a program, or a table.

I want to take advantage of advanced search. So I'm going to go ahead and click on the button that says Advanced Search. And now, I have a list of filters I can use to narrow down the result or tables that will appear. For this video, I am interested in median home values. So the first filter that I'm going to apply is I'm going to go to the topics filter. I'll go into Housing, because, again, it's median home values. And let's click on where it says Financial Characteristics.

We have several options here. I'm going to go with the checkbox for Housing Value and Purchase Price. Because again, median home values-- that sounds about right. There's other filters I can apply. I'm just going to do one more for this example here. I'm going to click on Years on the browse filters on the left hand side. And for this tutorial series, I will stick with 2018 data. So I'll select that checkbox. You'll notice at the bottom, as I select my filters, you see them applying at the bottom Selected Filters.

Now that I have those filters in place, I will click Search. And let's take a look at the results that we get. Right now, we are on the All results page. And here, I can choose to look at tables that are available, some maps, or pages that satisfy those filters that I just applied. I am interested in the actual table, so at the top of the page, I will click on Tables. And now, I see a list of tables on the left hand side that I can choose from.

For this example, I'm going to select on where it says Median Values (dollars). And the table ID that I'll be using here is B25077. When I selected that I can actually see that table appear up on the right hand side. And because I didn't apply any sort of geographic filter, I just see the statistic summarized at the national level. Particularly if you look at the Product dropdown menu, you see that this is for 2018. It is the One Year Estimates detail table.

So what do I mean by one year estimate? And where is this data actually coming from? Well, on data.census.gov, there's data from multiple surveys and censuses. And, in this course, we will use data which comes from the American Community Survey. In short, we can call it ACS. The ACS is the nation's largest ongoing household survey. And it provides social, economic, housing, and demographic data every year.

Now, the ACS is based on a sample. So, you'll notice here in the table that there is a margin of error that is provided to account for any of those sampling errors. In addition to that, they make available one year and five-year estimates. Five-year estimates are available for many geographical areas, all the way down to the block group level, while one-year estimates are only available for all areas that have a population of 65,000 or more and for metro or micropolitan statistical areas of 50,000 population or greater.

In this class, we will analyze data at the regional and state levels, and we're going to use the ACS one-year estimates. Again, I'm interested in the median home values at the state level, not the national level. So, in order to apply a couple more filters and customize this table, I'm going to click on the Customize Table button.

When I do that, I actually see this Advance Filter menu appears. And some of these filters are actually the same filters that we saw on Advanced Search. For example, I see a year filter applied for 2018 and a topic that we selected earlier as well. For me, I want to focus now on modifying the geography. So I'm going to click on Geographies on that Advanced Filters menu. I want state level. So I'll click State. And you can pick and choose the states that are of interest to you. I just want all the states, so I'll select the checkbox that says All States in United States. I'll click Close.

And now, you'll see that my table has been updated. I now see the different states represented in here. I have a column, for example, for Colorado and an estimate, as well as the margin of error. And I can just scroll through this table to see all the states being represented. I do want to make a note here. There is a chance that if you're following along with me, the states might appear in a different order, but you'll still have the same data available to you.

All right. I have the data that I'm interested in. I am now ready to download this data. So I'll click on where it says Download. I'll make sure that the one year estimates for 2018, that checkbox is selected, and I'll download this as a CSV file. Click Download. Once the files are prepared, I can click Download Now. And this is going to download a Zip file containing the data that I just selected. Let's take a look at that Zip file. I'll go ahead and open it up.

I'll expand this window a little bit so we can take a look what's here. You notice that there's actually three files contained in this Zip file that I just downloaded. There are two CSV, or comma separated values, files, as well as a text file in here. The one that we are going to be interested in is the CSV file that contains the word data_with_overlays in the title.

So I'm going to go ahead and double click on that. And this is just going to open up this file in Excel. I just want to take a look at it and make sure we got what we are interested in. So let's take a look here. I'm going to zoom in on this Excel file a little bit more, and I am also going to optimize the view of the columns just so we have enough room. A quick way to do that in Excel,

I'm going to, in the upper left hand corner, click on the Select All triangle. Then on the Home tab in the Cells group, I'll go to Format and then select out of Fit Column Width. I think just to make it a little bit easier to see the values here.

So let's take a look here. You notice I have a column called GEO_ID. I have IDs from my different states available. The name of the state, then I have the mean home value estimate-- again, the one year estimates for 2018. And then finally, in column D, I have that margin of error value as well. Now, at this point, I could actually take this file and import it into SAS Studio as is. And that is totally fine.

But in order to simplify the import process a little bit more and while I already have this file opened up in Excel, I'm going to make just a couple of changes. Again, these changes are not necessary. And you can even do these changes in SAS Studio. I'm just going to do them now-- again, to simplify that import process.

So first thing I'd like to change are the column names. There are some rules for column names that we have in SAS. So in SAS, column names must be one to 32 characters in length. It's recommended that the column name begins with a letter or an underscore, and that it continues on with letters, numbers, or underscores.

Now, in addition to those rules it would be nice if those column names are in the very first row of this file. So let's go ahead and change up these column names. I will leave, in Cell A1, GEO_ID, as is. But for cell B1, instead of Name, let's call this State. Cell C1-- I'm going to rename this to MedianHome-- one word. And then cell D1, I'll type in MedianHomeMOE-- again, one word. So that's change number one that I made.

Since I have those column names that I want in that very first row, I actually no longer need this second row that has these descriptive labels contained in there. So I'm going to go ahead and delete it. What I'll do is right click on the row number, 2. And then I'll select Delete. Last thing that I'm going to do while I have this file open in Excel is to apply currency formats to column C and D. Because, again, these are median home values, they represent currency values.

What's nice about formatting them in Excel is that when I import this file into SAS Studio, those currency formats are automatically going to be applied and come through over in SAS Studio. So to apply a format, I'm going to click on the column heading C, hold down my control key, and then also select column heading D. I can right click anywhere in this highlighted region, select Format Cells. And then we have the Format Cells window that pops up.

From the category, I'm going to go into the Currency category. Let's go ahead and decrease the decimal places to zero because there are no decimal places to start with. And I'll leave everything else at the default. I'll click OK. And here we have it. Now, we see column C and D applied as currency values.

That's all I need to do now, so I'm going to go ahead and save this file. I'll go to File, Save as. Feel free to navigate to a location of interest. I'm just going to go in and pick a location. I am going to name this file median home value. And I could keep this as a CSV file, but to make sure that the formats that I applied are properly saved and that they'll be used whenever I do the

importing into SAS Studio, in that Save as type drop down list, I'm going to instead choose Excel workbook with that .xls extension at the end. And click Save.

That's all I need to do. I have the file ready to go. I can go ahead and now close the Mean Home Value.xlsx file. Now you know how to access and download data from data.census.gov. Make sure to try out the practices in the tutorial notes. To learn something new, try out the challenge practice. It'll show you how you can extract just a subset of the data that you see when customizing the table in data.census.gov. Thanks for watching.

Import Census Bureau Data into SAS Studio | Video 3 of 6 | 13:41 minutes

Hi, everyone. Welcome to the Analyzing Census Data in SAS Studio Tutorial series. In the previous video, I used data.census.gov to download data on median home values in each state. I also prepared this data in Excel for import. So, in this video, let me show you how easy it is to import this data into SAS Studio using the Import Data utility. If you want to follow along with me, make sure you watch the introduction video to set up your environment and tutorial data. You can find a link to it below or on the main tutorial page. Let's get started.

So I've opened up SAS Studio, and I am using SAS Studio through On-Demand for Academics, which is free for learners. But, again, if you're using your own installation of SAS Studio. That is totally fine as well. Just to introduce SAS Studio a little bit here, SAS Studio is a browser-based, SAS programming interface. And it can connect to a local or hosted SAS server.

You can write your own code if you would like to do that or you can use the interface for these tasks that will generate SAS code for you automatically. To show you the SAS Studio interface a little bit, on the left hand side, we have our navigation pane. This is how we'll access all of our files. And on the right hand side, we have the work area. The work area is tab-based. So as we open up files or we try out a task, you'll see new tabs open up in that work area.

So now for importing. Before I can actually import that Median Home Value Excel workbook, I need to upload it to the SAS server where the processing will occur. So that's my first step. In the Server Files and folder section in the Navigation pane, I'm going to select the Census Data Analysis folder because that's where I want to upload my file.

I'll click on the Upload icon with that arrow pointing upwards. I'll select Choose Files. And I'm going to grab that Median Home Value Excel file I created in the previous video. Click Open, and then Upload.

So, now, if I go ahead and expand that Census Data Analysis folder, I should now see my Median Home Value Excel file. Now, if you didn't follow along with me in the previous video where we accessed data from data.census.gov and prepared it and Excel, no worries at all. In your Census Data Analysis folder, you should see the Median Home Value underscore s dot xlsx file, and you can use that file instead.

I'll, again, work with Median Home Value. To start the importing process, I will right-click the file name, and select Import Data. This is going to open up the Import Data utility on a new tab in the work area.

Now, you'll notice that the default view is split where I see the settings for the utility on the top and the code and the results that are generated on the bottom. I'm going to click on the Settings view just so I can focus on the settings for now. Now, I will mention that in this specific tutorial, I'm importing in an Excel file, and in order to use the Import Data Utility to import an Excel file, I need to have a license for SAS Access Interface to PC files. That is included with SAS On-Demand for Academics which, again, is free. We just want to make sure that you have that available to you.

So let's jump in with our settings here. The first part I see, I can see that we are importing in the Median Home Value Excel file, and I can type in a worksheet name. By default though, the import data utility is going to import in the first worksheet. Our Excel file only has one worksheet in it, so we don't need to type in the worksheet name here. I can skip over that.

Next, moving on to the output data section. Whenever we perform an import, essentially, what's happening is we have this Excel file, and the import is creating a copy of that Excel file as a SAS table. And I specify a name and location for that new SAS table that I'm creating. So I'll click Change.

For the library, where I want to save this, I'm going to select the Census library that we created. Remember this is a pointer to our Census Data Analysis folder. And in the data set box, I will call this file Median Home Value, and click Save.

Moving down to Options, this File type dropdown list. You can actually leave this at the default option because SAS can figure out what that file type is based on the extension of the file, but I'm going to go ahead and select it to be explicit. So using the dropdown this, I'll select XLSX because we have a Microsoft Excel workbook.

Finally, you'll notice this checkbox Generate SAS variable names is already selected by default. With this option selected, SAS will generate SAS column names from the data values in the very first row of the Excel file, which is why I went through the process in the previous video to give them proper SAS column names. And that is it.

Before I run this utility, I am going to change my view to Code/Results. And I'd like to show you that all the Code tab here, you can actually see the SAS code that was generated behind the scenes. I didn't have to type this myself. Just by selecting the options on the settings, SAS generated this for me.

Let's give it a run. I'm going to click on the running man icon, Run, or if you'd like, you can select F3 on your keyboard as an alternative. Taking a look at the Results tab, which is open here by default, this shows us the attributes of the new SAS table that was created. Take a look at the Data Set Name field. Notice the name of this table is census.medianhomevalue.

This two-level naming convention is used whenever we create tables and libraries. So we first have the library name census dot the name of the table, median home value. There's other pieces of information here, like the number of observations, or rows, 52, number of variables, or columns, for.

But I'm going to move on to the Output Data tab. So I can actually see what that data looks like. I see four columns here. I see the names of GEO_ID, State, MedianHome, and the MedianHomeMOE. And you might remember that these were in the very first row of that Excel file. And SAS automatically used those as the column names as we see here. I can see right above the table that, again, there are 52 rows and 4 columns.

And now I have those median home values along with the margin of errors for all states available to me as a SAS table in SAS Studio. Now, what I like to do is explore this table a little bit more.

But before I do that, I do want to point out one other tab we have, which is the Log tab. The log term is where we view the SAS log. And the log essentially displays messages that are returned back from SAS.

There were no errors or warnings, which is good. But we do have a couple of notes here. If you expand the notes section, you can click on these notes to jump to where that note occurred within the log itself. But there are no notes that are of concern in this example. Let's go back to the Output Data tab and explore this table a little bit more by modifying the view of this table.

So let's start with the columns area. You'll notice in the columns area, by default, all columns are selected. I just want to focus on a couple, so I will clear the check boxes for GEO_ID, and then also MedianHomeMOE. So now, I'm just viewing State and MedianHome.

If I want more information about a particular column, for example, its attributes, I can select the column in the columns area. So I'll select MedianHome. And on the bottom I see those attributes. For example, I can see that this is a numeric column. It actually has a format applied to it. It's the NLMNY15 dot format. You can think of this as the national language monetary format.

What this means is that, as a numeric column, behind the scenes, the stored value only contains the numbers and the decimal point, for example. But, by having this NL Money 15 dot format applied, we are actually viewing, or this value is being displayed with the local currency. Now, in our example, that means there will be a leading dollar sign. There will be commas separating each set of three digits. No decimal places, all within the total allotted width of 15 we see here.

And that's why we're able to see this value with the dollar signs just like a currency. Reason why this was automatically used for us was, remember, whenever I was in Excel, I applied a format in Excel. And whenever I did that and import the table in, an equivalent is automatically applied in SAS Studio, which is, again, the NL Money 15 dot format.

Let's go ahead and further modify the view of this table. Maybe I'm interested in only looking at the states with a median home value greater than $300,000. What I can do is right click the column I want to apply a filter on, MedianHome, and then select Add Filter. First, instead of equality, again, I'm interested in it being greater than $300,000. So I shall go with greater than, or equal to, select that symbol. And then type in 300000.

You want to make sure that you type this value without any dollar signs or commas because we want to match the unformatted stored value. Click filter. You see the view has been updated, and now I see that there are 11 states that satisfy that criteria.

Last thing that I'll do here is I'd like to go ahead and sort by the Median Home Values as well. So, like I did earlier with adding the filter, I'll right-click the Median Home Value column, the column heading again, and then I'll go with sort descending. So this now makes it easy to see that the state with the highest Median Home Value was Hawaii with roughly $631,000. And, again, there were 11 states that had Median Home Values greater than $300,000.

Now, I would like to mention here that any of these customizations that we're applying to this table in the table viewer, like sorting our rows or applying any sort of filter, these are not saved

with the table itself. However, what SAS is doing behind the scenes as we're selecting these options is it's actually generating SAS code that we can use. If you'd like to take a look at that code, you can click on this button right here that says display the code that creates the current table. When I click on that, a new program tab opens up, and it displays the code that was used to generate that specific view of the table. So if you want a starting point, you're able to use this code right here.

I'll go ahead and close this Program tab. Now, let's go ahead and remove those customizations that we applied and go back to the original view of the table. So, first thing I'll do, is next to where it says filter, and I see Median Home Value greater than or equal to 300,000. I'll click on the X, Clear Filter to remove the filter, and now I see all the rows.

Now, to sort it back in the original order, I will right click MedianHome, the column heading again. This time select Sort by Data Order, and if I want to bring back the columns, I can go back to the columns area and select those. But, instead, I'm going to click on the Refresh icon, and now I have all of my columns again.

At this point, I am done. So I'm going to go ahead and close the Median Home Value tab. And I am not going to save these changes. It's important to keep in mind that it is not necessary to save the settings that we specified in that import data utility to save the imported data. That imported data was actually saved just by running the utility and creating the output table Census dot MedianHome.

And I can actually take a look at that table. I can take a look at the library section in my navigation pane, expand My Libraries, Census, and there we have it. I can actually see the Median Home Value table is available to me. Remember, again though, that census library is just a pointer, or shortcut, to the Census Data Analysis folder. So if I go to Server Files and Folders, and look at my Census Data Analysis folder, I should see that same Median Home Value SAS table available in here as well.

You've seen how to import census data into SAS Studio. Now, it's your turn to try this out with the practices in the tutorial notes. If you want to take this process a step further, the challenge practice will show you how you can transpose, or restructure, the imported table with the Stack/Split Columns task. Thanks for watching.

Visualize Census Bureau Data in SAS Studio | Video 4 of 6 | 17:19 minutes

Hi, everyone. Welcome to the Analyzing Census Data in SAS Studio tutorial series. In the last two videos, I downloaded data on median home values in each state, then imported this data into SAS Studio. Now that I have a SAS table that I can use, I'll show you how you can create a bar chart using the bar chart task to have a visual to compare median home values across states. We'll also look at editing the code generated by the bar chart task to further enhance the results.

If you want to follow along with me, make sure you watch the introduction video to set up your environment and tutorial data. You can find a link to it below, or on the main tutorial page. Let's get started. So in SAS Studio, I have the census.medianhomevalue table opened up. We're just going to take a quick look at this table before we jump into the bar chart task.

Now if you didn't follow along with the previous tutorial where I imported in this data, no worries at all. You can instead use the census.medianhomevalue_s table and follow along using that. Looking at this table here, we have one row for each state, and we have this column MedianHome containing those median home values. What I want to do with the bar chart task is essentially provide a visual of this information, to make it easier to compare these median home values. Let's jump into the task.

In the navigation pane, I'm going to expand the Tasks and Utilities section. And under Tasks, I'm going to expand the Graph category, and then double click Bar Chart. This opens up the Bar Chart task in a new tab in the work area, and by default, the view is set to split. So I see the settings on one side and the code and results on the other side. I'm going to keep that as is, but I'm going to click on the Maximize View button, and that's going to hide the navigation pane, and just provide me with a little bit more room to work with.

So the first thing I need to do on the settings side, specifically under the Data tab, is specify my input table. Now you'll notice that my input table is that census.medianhomevalue is already selected, but if you don't see that, click on the Select a Table button, expand that Census Library, select the median home value table, or the underscore S version, and click OK.

On that Data tab, next, a chart orientation, I'll keep this as a vertical bar chart. So let's move on to our task roles. Our task roles determine how our columns will be used within the task. We have a couple of roles here, but take a look at the Category role. The Category role has a red asterisk next to it, which means that it is a required role. I have to assign some sort of column to this role.

Now remember, I want to create a bar chart with one bar for each state, the height representing that median home value. Because I want one bar for each state, to my Category role, I'll assign that State column. So I'll click on the plus sign, which is add a column, select State, and click OK.

Next, for my Measure role, that determines the height of the bars. The default is just a frequency count, but again, I want those median home values. So from the dropdown list I will select Variable, and to the Variable role I will assign the MedianHome column. And click OK.

Now for the statistic that is used, the default is the sum. And we're actually going to use that statistic because we only have one median home value per state, sum of one value is just the value itself, so again, we're going to stick with that Sum.

Now there is a lot more options that we can take advantage of in the Bar Chart task, and we'll explore some of those coming up. But for now, let's take a look at the Code tab. This task, the Bar Chart task generates PROC SGPLOT code behind the scenes for us. And we can see that right here on that Code tab.

At this point, let's go ahead and give this code a run, let's see what we have so far. So I'll click Run. Now I'm looking at the bar chart, we have a start here, but we could probably make some improvements. All right, I have one bar for each state, the height represents the median home value, that's good. But I can see that all of my state names are being overlapped with each other. I might want to expand the width of this graph to make sure that they aren't overlapping. Maybe I'd like to add in a title, and maybe I can change the label of the y-axis, for example.

To make these changes, I'm just going to go back to the settings side. Specifically now, we're going to the Appearance tab to make these changes. First, to change the label on my y-axis, MedianHome, I'm going to expand the Measure Axis heading, then from the Display label dropdown list, I will select Custom label. In the label box, less label this Median Home Value.

Next, the title. I'll expand the Title and Footnote heading, and let's call the Title Median Home Value in 2018. Finally, the graph size. I'll expand the Graph Size heading, and for the Width, let's bring that up to 10 inches, and I'll leave the Height at the default 4.8 inches. Let's run it again, let's see how our bar chart changes.

This is looking a lot better now. My state names are no longer overlapping, there's more room for it to work with. I see my Median Home Value label on my y-axis and a nice title as well. Now you might have noticed that when I ran it the second time, next to my Bar Chart tab, there's actually a warning icon. If you see an icon like this, it's always a good idea to take a look at your log. So let's jump into the Log tab.

I see there are two warnings here, and they say that the width exceeds available space for RTF, which is like the word format, and the PDF destination, setting width=8in. So, what does that mean? Well by default in SAS Studio, our results get generated in the HTML5, RTF and PDF results.

What these warnings are saying is that for the RTF and PDF format, specifically, that width of 10 was too wide, so it just set it to 8 automatically. This does not affect our results though because on the Results tab, we are only viewing the HTML5 results. But it's always good to make sure you check your log to see if there's any warnings or errors you might want to be aware about.

I'm going to go back to the Results tab, taking a look at my bar chart again. Now this time, what I'd like to do is maybe plot the US median home value as a reference line in this bar chart. I don't know what that US median home value is, and it's not a part of my data. But, data.census.gov makes it really, really easy to find that information.

So in a new tab, I'll go to data.census.gov. This is a great resource to access data that the Census Bureau makes available. And what I'll do this time is in the free form single search bar, I'm going to just type in exactly what I'm looking for, which is median home value 2018. Let's see what we get.

Now previously I was looking at tables, but I actually don't even need to look at a table to get the information that I'm looking for. You'll notice at the top of the All results page, it immediately tells me that the median housing value in 2018 was $229,700. So that was really easy for me to get. With that in mind, let's go back into SAS Studio. On the Appearance tab, let's add in that reference line.

Under the Measure Axis heading, which is already expanded, I'm going to check the check box for Create a reference line, and type in that value, 229,700. Let's add in a custom label as well, and I will label that as U.S. Median. Let's give it a run.

So I get pretty much the same bar chart, but now I have that U.S. median home value plotted as a reference line in green. This makes it pretty easy to see which states are above the U.S. median home value or below it. Now let's take that a step further. Instead of plotting the U.S. median home value as a reference line, maybe I want to filter my values based off of that, so only plot the states that have median home values exceeding the U.S. median home value.

To do this, first I'm going to clear that reference line. So I'll clear to create a reference line checkbox on that Appearance tab. To filter my data, I'm going to the Data tab. And right below my input table, I'll click Filter. In the Filter expression box, I will type in MedianHome --that's the name of the column-- is greater than 229,700.

The syntax that is used here in the Filter expression box is based off of the SAS SQL procedures WHERE clause, just without that WHERE keyword. If you're interested to learn more about that syntax, take a look at the tutorial notes, there is a link to the documentation there. But this is good to go for me so I will click Apply and let's run it again.

So in my updated Bar Chart, I now see less states. These are states that have median home values exceeding that $229,700. So my bar chart's looking pretty good so far, but you'll notice right now that my bars are sorted in ascending or alphabetical order by the states. Maybe instead I'd like to sort my bars by the median home value, instead of by the states. Well, let's take a look at Tasks, see if there's any options for that.

So let's take a look at the Appearance tab. And under the Category axis heading let's see what we have. So there's an option, Reverse tick values. What that'll do is sort to bars in reverse or descending order by the states. Not quite what we want.

And then there is Show tick values in data order. So sort the bars, or the states, in order in which it appeared in the input table itself. But there isn't an option that lets me sort my bars by those median home values. If there is an option or setting that you're looking for that you're not finding in the task itself, what you can do is go in and modify the code that was generated by this Bar Chart task.

If I go to the Code tab there is that code that we want to modify. Before I go in and do that, there's one more resource I'd like to show you. And that is on the Information tab. I mentioned earlier that the Bar Chart task generates PROC SGPLOT code.

And if you're not very familiar with it, you want to learn more about it, that information tab actually has a direct link to the documentation that talks about the SGPLOT procedure. So you might want to check that out. But for now, let's go back into the Code tab.

Now the Code tab, this is a read only copy of the code. If I want to make a modifiable copy I can just click on Edit. This opens up a new Program tab that is again, modifiable, so I can make some edits to the code that was generated. So you might not be familiar with SAS code, but SAS Studio actually has a lot of features available that make it a little bit easier for you to code. And has some features to help guide you.

So I'd like to show you a couple of those features. First, let's say in my code I'm not very familiar with this VBAR statement, but I want to learn more about it. What you can do, go ahead and right click on the VBAR keyword and select Syntax Help.

The Syntax Help window tells me a little bit about this VBAR statement. So the description says, it creates a vertical bar chart that summarizes the values of a category variable. And I can also see the syntax. After that VBAR keyword I list a category variable. And then after a forward slash is where all of my options are going to go.

So you might be wondering, well, what are some options that are available? Is there something that I could take advantage to sort my bars by the median home values? Well, let's see.

I'm going to place my cursor right before the semicolon on that VBAR statement. You just want to make sure you're after that forward slash. Because again, we're interested in those options.

When I hit the space bar an autocomplete window opens up and shows me all the valid options for a VBAR statement. And there are a lot of options I can take advantage of here. The one that I'm interested in is the CATEGORYORDER= option. So I'm going to hit the C on my keyboard, takes me right to the C's and now you see CATEGORYORDER= is highlighted.

Now I have the Syntax Help window for the CATEGORYORDER= option. I see that it specifies the order in which the response values, our median home values, are arranged. That's exactly what we want.

So with the CATEGORYORDER= option highlighted in my autocomplete window, I can just hit the Enter key and now the CATEGORYORDER= option is automatically added into my program. Which is nice. Then autocomplete opens up again.

Now I can see the valid option values for the CATEGORYORDER= option. RESPASC and RESPDESC. So response ascending or response descending.

I'm going to single click on response descending so I can read more about it. As you can probably guess, it sorts the response values in descending order. That's exactly what I want. So this time I will double click it, and that also adds in that value to my program.

So the Syntax Help window and Autocomplete window can guide you out and help you out as you're typing inside of a SAS program. There's one more thing I'd like to add in here, and that is a secondary title explaining the filter that was applied.

So right below the TITLE statement I'm going to add a TITLE2 statement, and in quotes, type in the title. I'll say "States Exceeding the US Median of $229,700". And then end that statement with a semicolon. Let's give the program a run.

So here is my final bar chart. Again, only looking at the states that exceed that US median home value of $229,700. And now it is sorted in descending order by those median home values. At this point, you may want to share this bar chart in let's say a PDF format for example.

Earlier I mentioned that by default SAS Studio generates our results in the HTML5, PDF, and RTF formats. In order to access those, there are these buttons on the results tab to download the results as an HTML file, PDF file, and an RTF or a Word file.

As an example let me show you the PDF file. Just click on the button, open that up, and now I have the same exact bar chart but now in a PDF file. Which may make it a little bit easier to share with your colleagues. I'll go ahead and close out of my PDF file. And let's save our work.

I'm going to first exit out of maximized view, bring back my navigation pane, and to save the program on the code tab I'll click on the Save icon. Let's save it in the Census Data Analysis folder. I will call this program Median Home Bar Sort. And click Save.

Closing out of the program. Let's also save the settings specified in the bar chart task. Click the save icon here as well. And in that Census Data Analysis folder let's call this one Median Home Bar, and click Save.

Now you know how to create a bar chart using the Bar Chart task to visualize Census data. Make sure to try out the practices in the tutorial notes. To learn how to create a stacked bar chart, check out the challenge practice. You'll even edit the code generated by the bar chart task to create a stacked bar chart, where each bar equals 100%. Thanks for watching.

Prepare Census Bureau Data in SAS Studio | Video 5 of 6 | 16:45 minutes

Hi, everyone. Welcome to the Analyzing Census Data in SAS Studio tutorial series. In the previous video, I created a bar chart comparing median home values across states, and I was able to do this with minimal preparations needed with the data. However, sometimes you may want to prepare or inform your data before creating a visualization or performing an analysis.

For example, you may need to combine multiple tables together, filter your data, or create new columns. In this video, we'll create a listing report that displays the top five states by median household income within each geographical region. To create this report, we'll need to prepare our data.

First, we'll use the Query utility to combine median household income data with a geography lookup table. Then we'll use a Rank Data task to rank the median household income values within each region. Finally, we'll use the List Data task to create the report.

If you want to follow along with me, make sure you watch the introduction video to set up your environment and tutorial data. You can find a link to it below or on the main tutorial page. Before we get started with our SAS Studio task, I'd like to quickly show you where the data is coming from.

For our median household income values, you can find this information on data.census.gov. The table ID is B19013, and we'll be working with the data from 2018, specifically, the one-year estimates for all states in the United States. In addition, we have a geography lookup table that could be found on the Census Bureau's website as an Excel file. I've already downloaded this file just to show you what it looks like. But this explains which region and division each state belongs to.

If you'd like to learn more about census regions and divisions, the Census Bureau makes a PDF available explaining what those are. Now, the median household income values table, as well as the geography lookup table, have already gone through the process outlined in the Accessing Census Data and Importing Census Data videos. Those tables are already available to you in your Census library. So let's take a look at them in SAS Studio. In SAS Studio, in the Navigation pane, I'll expand the Library section, expand my Libraries, and then Census.

Our median household income values is in the MEDIANINCOME table. So I'll double click on that. And as we saw in data.census.gov, this contains each state's median household income value as well as the margin of error. In addition, the geography lookup table can be found in the GEO_LOOKUP table also in the Census library. So I'll double click on that.

And this table contains the regions, divisions, states, and associated codes. To view these tables together, I'm going to go up to the CENSUS.GEO_LOOKUP tab and drag that toward the bottom of the SAS Studio work area until a green checkmark appears.

When that appears, you can let go of your mouse, and now we have a stacked view of the tables. Now, in order to rank the median household income values within each region, these two tables need to be combined or joined together. A join will allow us to take two or more tables and

combine them horizontally on one or more common columns. This will allow us to select data from multiple tables as if the data were in one single table.

Now, looking at these two tables, it seems that the State column is the common column. Not necessarily because they have the same column name of state but more so that they contain the same information, which are already full state names. In order to perform a join to combine these two tables together, we can use a query.

A query will allow us to extract data from one or more tables according to criteria that we specify. Before we start a query, I'll first close the CENSUS.GEO_LOOKUP as well as the CENSUS.MEDIANINCOME tabs. To start a new query, on the toolbar I'll select the New Options button and then select New Query.

As always, the default view is split where we're seeing both the query settings as well as the code and results together. I'm going to change this view to just settings. The first step to working with the query is to add a table to the query. So from the library section, again, in that Census library, I'm going to drag the MEDIANINCOME table onto the Tables tab to add it to our query.

To join the geography lookup information with the median household income values from our Census library, this time I'll drag the GEO_LOOKUP table and drag it to on top of the MEIDANINCOME table. A join is automatically created if the tables include columns with matching names and data types. So in our case, a join was automatically performed on the state columns.

You'll notice that the default join type is an inner join, which returns only to the subset of rows from the first table, MEDIANINCOME, that matches rows from the second table, GEO_LOOKUP. In other words, only states that are found in both tables will be included in our final output table. As a side note, we drag and drop tables to add tables to the query as well as to perform a join, but alternatively, you can click on the Add button on the Tables tab to add a table or to add a join. And you can also modify the join conditions as well.

With the tables now joined together, let's specify what will be included in our output table using the Columns tab. First, to include all columns for the median income table from the columns lists, I'll drag the median income table onto the Select tab. Back in the Columns list from the geo lookup tables, I'll expand GEO_LOOKUP. I'll drag over the Region as well as the Division column.

Alternatively, you can use the Select Column button to add columns to the Select tab as well. I'm going to use the Move Row Up and Move Row Down buttons to rearrange these columns. So I like them in order of GEO_ID, State, followed by Division, Region, MedianIncome, and then MedianIncome_MOE.

Next, let's move onto the Sort tab. First, I'll drag Region onto the Sort tab, and I'll leave the sort direction as Ascending. Next, in the Columns lists, I'll expand the MEDIANINCOME table and drag over the MedianIncome column to the Sort tab. For the MedianIncome column, in the Sort box, I'll change the sort direction to Descending.

What this means is that the table will first be sorted by Region in ascending order, and within each region value, the rows will be sorted by MedianIncome in descending order. Now, it's important to keep in mind that it's not necessary for us to sort our data before we use it in the Rank Data task to rank our data.

However, by sorting our data by groups and the value that we'll be ranking on, which is MedianIncome, the output table that gets produced by the Rank Data task will be sorted in rank order. So that's why we're doing it here. The last thing I'll do is change the output table name. So on the Properties tab, I'll just first quickly verify that the output type is table.

For the output location, I'll type in census for the Census library, and in the output name box, I'll type in medianincome_geo. That'll be the name of our output table. Before I give this a run, I'm going to change the view to Code/Results, and here you can see that the Query utility generates structured query language or SQL code. All right, let's give this a run.

Taking a look at our output table, you can see that this table combines the median household income data with the geography information. Now that we have the median household income values and geography information combined, we can now rank the median household income values within each region. To do this, I'm going to use the Rank Data task. To start the task in the Navigation pane, I'm going to expand the Tasks and Utility section, expand Tasks, and then under the Data category, you'll find the Rank Data task. So I'll go ahead and double click on that.

Again, the default view is split, but I'll click on Settings just so we can focus on the task settings first. First, on the Data tab for input table, I'm going to click on the Select a table button. And from the Census library, I'm going to select the table we just created, MEDIANINCOME_GEO, and click OK. For our task role, each column that is assigned to the columns to rank role will be ranked.

We want to rank our median household income value. So I'll click Add Columns, select MedianIncome, and click OK. Let's take a look at the other roles that are available. I'll expand Additional Roles. Now, when you assign a column to the Rank by role, rankings will be calculated within each group. We want rankings within each region, so I'll click Add Columns, select Region, and click OK.

Let's modify the name of the output table. In the Data Set Name box, I'll type in census.medianincome_rank. You'll notice this option, Create New Variables for the Rank Variables. I'm going to make sure that is selected. This specifies that the output table that we are creating will contain the original column as well as a ranked column.

Moving on to the Options tab, I'm going to leave the Ranking method of Ranks, and If values are tied, use option Default Method, I'm just going to leave those as is. But I will change the Rank order option. Using the dropdown list, I'll select Largest to smallest. This means that a rank of 1 corresponds to the largest value in the group.

Let's change our view to Code/Results, and let's give this task a run. This table looks similar to what we had before, but now there is an additional column called rank_Medianincome. As you can see, the median household income values are ranked within each region from largest to

smallest. Note that this table is in rank order only because we sorted the input table by median income within each region.

The last step is to create a listing report that displays only the top five states within each region. To do this, we'll use the List Data task. So in the Task and Utilities section, I get in the Data category as well. I'm going to double click on List Data. I'm going to leave the view as split, but I'm going to click on the maximize view button to hide the Navigation pane. First step on the Data tab is to select our input table. So I'll click on to Select a Table button. And from the Census library, I'll select the MEDIAINCOME_RANK table.

Now, remember, I'm only interested in the top five states within each region. So I need to apply a filter to my data. Right below my table, I'm going to click on Filter. And in the filter expression box, I'll type rank_Medianincome less than or equal to 5. So only those states with a rank value of 5 or lower will be in our output report. I'll click Apply.

Taking a look at our task roles, columns that are assigned to the List variables role are printed in the report in the order that they are listed in this task role here. So to assign columns to this role, I'll click Add Columns, I'll select State, hold down the Control key and then also select median income, and click OK. Next, the Group analysis by role creates a separate listing for each distinct value of the column we assign to this role.

Because we want a separate listing for each region, I'll click Add Columns and assign Region to this role. Let's give this a run. Taking a look at our report, you'll notice that we do have separate reports for each of the four regions, and only the top five states are listed. You'll notice that instead of the column names, the column labels are printed, which are descriptive column headings. I'd like to enhance this report by removing this Obs column and replacing it with Region.

So back onto the Data tab, we can use the identifying label role to replace the Obs column with another column from our table. So I'll click Add Columns, select Region, and click OK. Let's run this again. And here is our updated report. You can see that instead of the Obs values, we now see the region instead.

The last thing that I like to do is modify the title for this report, which can be done by modifying the code that this task generated. So, on the Code tab, I'll click on the Edit button, and this creates a modifiable copy of the task-generated code. The only change I need to make is to the first line here, which is our Title1 statement. Inside of the quotes, I'm going to replace the current content with "Top 5 Median Household Incomes by Region." Let's give this a run.

Here is our final report with our updated title. Again, we see separate reports for each of the regions, and it makes it very easy to see the top five states. For example, in the Midwest region, we see that the state with the highest median household income value is Minnesota, followed by Illinois, North Dakota, Wisconsin, and then Iowa. I'll exit out of maximized view, and let's go ahead and save this program. On the Code tab, I'll click on the Save Program button.

I like to save this in the Census Data Analysis folder, so I'll select that folder. And I'll name this program Median Income Report, and click Save. I'll go ahead and close the Median Income

Report.sas tab and close the rest of the tabs as well, the List Data, Rank Data, as well as the query tabs. Remember that it's not necessary to save the settings specified in the Rank Data task or the Query utility because by running the task, the output tables are saved in the Census library, which, again, points to our Census Data Analysis folder.

You've seen how to prepare census data in SAS Studio to create a listing report. We specifically looked at joining tables together and ranking data within groups. Now it's your turn to try this out with the practices in the tutorial notes. If you want to learn about other data preparation techniques, such as creating a new column, and grouping values into categories, make sure to take a look at the challenge practice. Thanks for watching.

Analyze Census Bureau Data in SAS Studio | Video 6 of 6 | 20:40 minutes

Hi, everyone. Welcome to the Analyzing Census Data in SAS Studio tutorial series. In the last video, I created a listing report that displayed the top five states by median household income within each geographical region. In this video, I'd like to further analyze median household income values using some of the statistical tasks available in SAS Studio.

First, I'll show you how to use the Distribution Analysis task to examine the distribution of median household income values in each region. To determine if there is a significant relationship between median household income values and region, we'll use the One-Way ANOVA task. Finally, you'll see the Correlation Analysis task to examine the relationship between median household income values and other statistics, like mean hours worked per week and median monthly housing costs.

For this tutorial, I will assume some knowledge of statistical tasks and concepts. The focus will be more on how you can perform these statistical tasks in SAS Studio and how to interpret the results. If you'd like to learn more about the concepts mentioned in this tutorial, you can take our free e-learning for the Introductory Statistics One Introduction to ANOVA, Regression, and Logistic Regression course.

If you want to follow along with me, make sure you watch your introduction video to set up your environment and tutorial data. As always, you can find a link to it below or on the main tutorial page. Let's jump right in.

Let's start by taking a look at the data that we'll be using. In the Census Data Analysis folder, I'm going to double click on the stateinfo_combined SAS table. Along with some geography information in this table, for each state, we have information about the median household income value, the mean hours worked per week, the total population, the median age, the median duration of their current marriage, as well as the median monthly housing costs.

Now all of these estimates can be found on data.census.gov. In the table, you'll see the one year estimate for 2018 for all states. In addition, the geography information can be found in a geography lookup table that the Census Bureau makes available in an Excel file. I've downloaded this Excel file just to quickly show you what it looks like. But this file explains which region and division each state belongs to.

If you'd like to learn more about census regions and divisions, the Census Bureau does make a PDF available explaining those. You can find the links to all of the sites that I have mentioned in the tutorial notes, which can be found in the description below or on the main tutorial page. To combine the estimate as well as geography information into a single table, you can follow a process that's similar to what's outlined in the previous video on preparing census data.

However, this was already done for you. And again, all contained in the stateinfo_combined table. Using this table, I'd first like to examine the distribution of median household income values. To do this, I'll use the distribution analysis task. So in the Navigation pane, I'll expand the Tasks and Utility section. Expand Tasks. And from the Statistics category, I will double click on Distribution Analysis.

Keep in mind that all of the tasks that we mention in this video do require the SAS/STAT product to be licensed. If you are using SAS Studio through SAS OnDemand for Academics, then SAS/STAT is included. I'll start by clicking on Maximize View to go ahead and hide the Navigation pane.

First, to select our input table on the DATA tab, I'll click on Select a Table. And from the Census library I'll select the stateinfo_combined table and click OK. Now, the Analysis Variables role specifies the columns whose distribution we want to analyze. Since we want to analyze the distribution of median household income values, I'll click Add Columns, select the MedianIncome column and click OK.

Moving on to the Options tab, I would like a histogram. So I'm just going to quickly verify that that option is selected. And to add some density curves to our plot, I will select the Add Normal Curve and the Add Kernel Density Estimate checkboxes.

I'll also select the Add Inset Statistics checkbox so I can have some statistics included in my histogram. So I'll expand the Inset Statistic subheading. And let's include the number of observations, which are selected by default. The mean, as well as the median. All right, let's see what we have. I'll give this a run.

Taking a look at our histogram, the normal curve, which is in blue, outlined a normal distribution with the same mean and variance as our input data, stateinfo_combined. The kernel density curve, which is outlined in red, outlines a smooth approximation of the distribution of our observed data. You'll notice that these two curves appear to be fairly similar, although the data is slightly right skewed.

Taking a look at our statistics, you'll notice that the mean, which is roughly $62,000, is about $2,000 higher than the median, which is just under $60,000. Now, I'd like to know if the distribution is different across different geographical regions. So going back to our Options tab, I'm going to take advantage of the Classification Variables role. This can be used to create separate histograms for each classification level.

Because I'm interested to see if the distribution is different and each region, I'll click Add Columns and assign Region to this role. Let's give it a run and take a look at the updated results. You'll now notice that we have seperate histograms for each region. And each region does have a unique distribution.

You'll notice that the Midwest region has a right skewed distribution. The Northeast region seems to be closer to a uniform distribution. And the South and West regions have states with extreme values. So using the Distribution Analysis task, I can visually compare the distribution of these median household income values. And they do seem to be quite different among the different regions.

But now, I'd like to know if the relationship between median household income and region is significant. To investigate this relationship, I will use you One-Way ANOVA task. So first, let me go ahead and exit out of maximized view. And in the Task and Utility section, under Task, this time I'll expand the Linear Model section and double click One-Way ANOVA.

The One-Way ANOVA task will test and provide graphs for differences among the means of a single categorical variable, which in our case is our regions, on a single continuous dependent variable, which in our case, are the median household income values. I'll click on Maximize View again to high the Navigation pane.

And on the Data tab, you'll notice that census.stateinfo_combined is already listed as our input table. What happens is that the most recently used table in a task is listed as the input table by default. So we're good to go there. However, if you do need to select the table, you can always click on Select a Table to do that.

Taking a look at our task roles, the Dependent Variable role specifies a continuous numeric column. So I'll go ahead and click on the Add a Column button. And I'll assign MedianIncome to this role. Next, the Categorical Variable role specifies a column with values that specify the levels of our groups. So I'll click Add a Column again here. And I'll assign Region to this role.

Moving on to our Options tab, by default, the One-Way ANOVA test performs Levene's test for homogeneity of variance to test that the variances within each region are equal. This is one of the assumptions that we do need to satisfy for One-Way ANOVA. But I'll go ahead and clear the checkbox for Welch's variance-weighted ANOVA. If the ANOVA assumption of equal error variances across all groups is not met, then you can select this checkbox.

Moving on to the Comparisons method, by default, the task will determine whether there are significant differences in the mean of the median income between each pair of regions with Tukey's adjustment, so I'll leave that as it. And moving on to Plots, from the Display Plots dropdown list, I'll select Selected Plots.

I'll go ahead and clear the checkboxes for Means plot and LS-mean difference plot. And select checkbox for our Diagnostics plot. So we just want to check boxes for Box plot and Diagnostics plot. All right, let's go ahead and give this a run.

There are many parts to this output. I'm just going to point out a couple of portions. Taking a look at the overall analysis of variance table, you'll notice that the p-value is 0.0228, which is less than 0.05. What this means is that the test is significant. It suggests that there are at least two regions where the mean of the median household income values are significantly different. We'll see which regions are significantly different a little bit later on in the output.

Moving on to the Fit Diagnostics panel, this contains a set of graphs that are commonly used to help validate the assumptions of ANOVA. And there are several assumptions that we do need to make sure are met. The first assumption is that the observations must be independent. We don't see a period of repeated measures or clustering. And we can assume that the data was collected in a way to ensure independence of the observations. So the first assumption is met.

The next assumption is that the errors are normally distributed. The first scatter plot on the second row of the Fit Diagnostics plot is a quantile-quantile, or Q-Q plot. You'll notice that there is very little deviation from the reference line. So the residuals are normally distributed.

We can also verify this using the first histogram on the third row of the Fit Diagnostics panel. This is our residual histogram. And this histogram displays a relatively normal distribution of the residual. So again, satisfying our second assumption.

The third and last assumption is that all groups have equal error variances. And this can be verified with Levene's test for homogeneity of variance, which is a little bit later on in the output. I'll scroll down to there.

Notice that Levene's test returns a p-value of 0.1542, which is greater than 0.05. Therefore, the null hypothesis of equal variances fails to be rejected. So the assumption of equal error variances across all groups is met. So all of our assumptions for One-way ANOVA are met.

Scrolling back up to our box plots, you'll notice that the box indicates that the North East region has the highest average regional median household income. And the South region has the lowest. You can hover over any of the boxes or whiskers to display a tooltip that has some more descriptive statistics, if you're interested in that.

The big question, though, is which regions have statistically significant differences in the mean? So I'm going to scroll down to the Least Squares Mean table. In the Least Squares Mean table, first, you'll notice that each region is assigned a number. So Midwest is one, Northeast, two, South, three. And then West, four.

Taking a look at the pairwise comparisons among all regions, notice he adjusted P value of 0.0286 between regions two and three, which correspond to the Northeast and South regions. With this value being less than 0.05, it indicates that the mean of the median household income values between those two regions are significantly different.

The last thing that I'd like to do is determine whether there is any sort of relationship or correlation between median household income values and the other estimates and statistics that were included, like the mean hours worked per week or the median monthly housing costs. To do this, I'll use the Correlation Analysis task.

So first, I'll exit out of Maximize View, bring back the Navigation pane. And in the Task and Utility section, under Tasks, this is under the Statistics category. I'll double click on Correlation Analysis.

Clicking on Maximize View again to hide the Navigate pane. Again, notice that our input table is already listed as CENSUS.STATEINFO_COMBINED. So that's good to go. Moving on to our task roles, the Analysis Variables role lists the columns for which to compute correlation coefficients. So I'll go ahead and click on Add Columns.

And I want to assign everything aside from MedianIncome. A quick trick is to first select MeanHoursWorked, hold down the Shift key, and then select MedianMonthlyHousingCosts. This selects those two columns, as well as all the columns that are in between. And click OK.

Next, the Correlate With role lists the columns with which the correlations of the Analysis Variables are to be computed. Because I want to determine the relationship between median

household income values and the Analysis Variables, I'll click Add Columns and assign MedianIncome to this role.

Moving on to the Options tab. Under the Plots heading, I'm going to use the Type of plot dropdown list. And I'll select individual scatter plots. These individual scatter plots will display the pairwise relationship between our analysis variables and median household income. Let's give this a run.

Taking a look at our results, the Pearson Correlation Coefficients table displays the strength of a linear association between median household income and the potential predictor variables. The stronger the association between these two variables, the closer the Pearson Correlation Coefficient value will be to either negative 1 or 1, depending on whether that relationship is negative or positive respectively.

You'll notice that the Pearson Correlation Coefficient value between median household income values and median monthly housing costs is 0.93492. This indicates a strong positive linear relationship between those two variables as compared to the other pairs. In other words, the higher the median monthly housing costs, the higher the median income. We'll use the scatter plot to make sure and verify that that relationship is linear.

So let's take a look at a couple of those scatter plots that were generated. The scatter plots will provide a visual of the relationship between median household income values and the Analysis Variables that we selected. First, taking a look at the scatter plot for a median household income and mean hours worked per week, there doesn't seem to be any significant relationship between those two.

Moving on to the scatter plot for a median household income values and median duration of current marriage, there does seem to be a slight negative relationship between the two. This could indicate that the longer a couple has been married, the less the median household income value. However, there is an outlier here that may have had a strong influence on this relationship.

If you hover over this outlier, we can learn a little bit more about it. And we see that this corresponds to observation number 34. Based on the tooltip right now, I don't know which state this is. But we could look at the input table and locate observation number 34 to figure out which state it is. However, we will enhance the results to also include the state information in the tooltip.

Before we do that, let's look at one last scatter plot. And this is a scatter plot for median household income values and median monthly housing costs. The Pearson Correlation Coefficient value of 0.9349 indicated a strong positive relationship between these two variables. But now, looking at the scatter plot, we can verify that the relationship is indeed linear.

Let's now enhance our results by adding more information to the tooltip set up here in the scatterplot. And also, by sorting the Pearson Correlation Coefficients table. To do this, I need to modify the task generated code. So on the CODE tab, I'll click Edit to open up a modifiable copy of the code.

First, on the PROC CORR statement, right before the semicolon that ends the statement, I'm going to type a blank space. And this displays the autocomplete window with some valid options for the of PROC CORR statement. I'm going to type R to highlight the Rank option. And in the Syntax Help window, you'll notice that this option displays the ordered correlation coefficients for each variable.

I do want to include this option, so I'll go ahead and hit the Enter key to include the Rank option. Next, in my PROC CORR step, I can add in an ID statement that'll specify additional variables to include in the tooltip that appears in the scatter plot. So after the WITH statement and right before the RUN statement, I'm going to add in an ID statement and include the State and Region columns. And then end the statement with a semicolon. Let's give it a run.

Taking a look at our results, you'll now notice that the Pearson Correlation Coefficients table now displays the values in order from strongest to weakest in absolute value. And now, let's take a look at our scatter plots. I'm going to scroll down to the scatter plot for median household income values and median duration of current marriage.

Hovering over the outlier, we can now see that observation number 34 corresponds to District of Columbia. DC has a median household income value of roughly $85,000 with a median duration of current marriage at 10.7 years. Let's go ahead and save this program.

Back on the Code tab, I'll click on the Save Program icon. And in the Census Data Analysis folder, I'll save this program as Median Income Correlation, and click Save. I'll go ahead and close the Median Income Correlation tab and exit out of Maximize View.

You can go ahead and close out of the correlation analysis, One-Way ANOVA, Distribution Analysis, and stateinfo_combined tabs without saving the results. Now you know how to use several statistical tests in SAS Studio to analyze census data. There are so many more statistical tests available.

If you like to learn about another one, try out the challenge practice in the tutorial notes. It'll walk you through the linear regression task to predict median household income values based on the mean hours worked per week. To learn about the other statistical tests, take our free e-learning for the Introductory Statistics One Introduction to ANOVA, Regression, and Logistic Regression course. You can find a link below a description or on the main tutorial page. Thanks for watching.